TheDataOpen

Presented by Citadel and Citadel Securities
In Partnership with CorrelationOne

Problem Statement

Welcome to the 2021 Citadel Asia-Pacific Regional Datathon! This document explains the topic of the Datathon, important details about the datasets you'll be using, and guidance on how to submit your results.

Background

England is largely credited as the birthplace of modern football. Throughout the 18th and 19th centuries, the sport quickly proliferated throughout Europe, becoming by most metrics the most popular sport on the continent. The International Association of Football (FIFA) was formed in 1904 in order to centralize the sport's governance and overall management, and it has grown to be the effective steward of the sport globally.

Modern football is played with 22 players on the field at any given time, 11 players per team. Each team is afforded one player, called the goalkeeper, capable of using their hands to prevent the ball from entering the team's net. Defenders play directly in front of the goalkeeper and are tasked with stopping opposing players advances. Midfielders play one level up field from the defenders and typically play an amorphous role, shifting between attacking and defending based on the game's flow. Finally, forwards play closest to the opposing team's goal and are tasked with trying to score points for their team.[1]

Association (Club) football is widely considered to be the most popular sport across all of Europe. Within Europe alone, there are 37 professional football leagues spanning 30 countries and accounting for well over 1,000 clubs. Though this rapid proliferation is emblematic of football's popularity growth, many globally typically focus their attention on "the big 5" – five European leagues credited with the greatest talent and most storied clubs in the world. Out of the top 20 clubs measured by total value, 95% are members of a "big 5" league, and all reside in Europe.[2]

European football also extends further than a game – it is an independent economy, and a rather large one. The 4 most valuable franchises in the world are currently each valued in excess of $3 billion USD. In addition to franchise value, clubs typically spend excessive

[1] https://en.wikipedia.org/wiki/Sport_in_Europe#Association_football
[2] https://www.marca.com/en/football/internationalfootball/2021/01/26/601052cde2704ee7a58b45c8.html

amounts in order to acquire the talents of a particular player. In some cases, such as Neymar's transfer from Barcelona FC to Paris Saint-Germain, the fee paid to acquire a player can exceed €100 million EUR. In the end, rising ticket prices and exploding popularity make the proposition of managing a successful football club a remarkably lucrative venture capable of producing hundreds of millions of euros in revenue per year.[3,4]

Sports betting, largely through an intense wave of legalization and application development, has seen expansive adoption globally within the past 5 years. Mordor Intelligence predicts that adoption will only increase in velocity in the coming years, projecting a cumulative average growth rate of 10.1% in the market from 2019 to 2024. Further, Europe has been identified as the fastest growing sports betting market with football betting driving a strong majority of the growth. The total football sports betting market in Italy alone was valued at over €9 billion EUR in 2019.[5]

Your Task

Your goal is to use the various football data streams in order to discover and analyze patterns in football team performance and match outcomes, and make recommendations about optimal club construction or effective betting strategies. More broadly, you should analyze the given European football data holistically to identify the characteristics of clubs that make them most successful.

We have partially pre-cleaned several datasets for your use, including club playstyle attributes for nearly three hundred European football clubs, player skill attributes for roughly eleven thousand players, and match outcomes and betting odds for over twenty thousand matches from 2008 – 2016.

If you would like to enhance your analysis, feel free to use any other dataset you may find – just follow the guidelines in the "Additional Datasets" section below and be aware that the quality of your analysis will also be judged by the reliability of the data being used.

You are asked to pose your own question and answer it using the available datasets as well as any supplementary datasets that you may find. What is important is both the creativity of your question and the quality of your data analysis. **You need not be comprehensive; depth of insight is more important over breadth of the question posed.**

Submissions may be predictive, use machine learning, and/or time series analysis to predict outcomes of future matches. Submissions may also be illuminating, through the use of data visualizations or through sound statistical tests.

---

[3] https://www.bbc.com/sport/football/40762417

[4] https://www.forbes.com/sites/steveprice/2020/05/29/even-cristiano-ronaldo-cant-keep-juventus-in-the-top-10-most-valuable-clubs/?sh=2f821555e7f2

[5] https://www.mordorintelligence.com/industry-reports/online-sports-betting-market

Consider exploring one of the sample questions below, or creating your own variation. Creativity in formulating your own question is encouraged.; **however, it should not be at the expense of analytical depth, precision, and rigor, which are far more important.**

Sample Question 1: What mix of components (player skills, home/away match, team playstyle) is most likely to yield consistent wins?

Sample Question 2: Is there a dominant team playstyle? Is this playstyle capable of making up for a talent deficiency?

Sample Question 3: Is it possible to safely and effectively outperform an average market return by predicting and betting on match winners without running out of capital?

Datasets

The provided datasets are stored in the "Datathon Materials" folder on Google Drive. Your team should only use the datasets that are relevant to your chosen question/topic. The raw data sources are noted but we encourage you to use our data since they have been organized and pre-cleaned.

country
Reference dataset for 11 European countries with football leagues.
11 rows & 2 columns. Size: <1MB. Source: not public.

player
European footballers spanning 11 countries and 11 leagues, from 2008 – 2016.
11,060 rows & 5 columns. Size: <1MB. Source.

league
Reference dataset for 11 European football leagues across Europe.
11 rows & 3 columns. Size: <1MB. Source.

player_attributes
Skill attributes for ~11,000 footballers spanning over 35 categories.
~184,000 rows & 40 columns. Size: ~26MB. Source.

team
Reference dataset for ~300 European football clubs.
299 rows & 3 columns. Size: <1MB. Source.

team_attributes
Club playstyle attributes for ~1400 European clubs spanning over 20 categories.

1,458 rows & 23 columns. Size: <1MB. Source.

match
Match details for ~26,000 European soccer matches from 2008-2016. This dataset includes the match score, starting lineups, and gambling odds.
25,979 rows & 62 columns. Size: ~8MB. Source.

Additional Datasets

Participants are welcome to scour the Web for their own custom datasets to supplement their analysis. All additional data should be public and should not exceed 2GB zipped (consult the technical team if you believe your idea is worthy of an exception).

Some good starting points are this compilation of FIFA game data, as well as this repository of historical betting data and match outcomes for European football. You may also want to draw on open source football data to enrich the given datasets.

Other Materials

We will provide you the schema for each of the data tables in another packet.

Submissions: Content

Submissions should have two components:

1. Report – this should have two main sections:
    a. Non-Technical Executive Summary – What is the question that your team set out to answer? What were your key findings, and what is their significance? You must communicate your insights clearly – summary statistics and visualizations are encouraged if they help explain your thoughts.
    b. Technical Exposition – What was your methodology/approach towards answering the questions? Describe your data manipulation and exploration process, as well as your analytical and modeling steps. Again, the use of visualizations is highly encouraged when appropriate.
2. Code – please include all relevant code that was used to generate your results. **Although your code will not be graded, you MUST include it or your entire submission will be discarded.**

Additional information (e.g. roadblocks encountered, caveats, future research areas, and unsuccessful analysis pathways) may be placed in an appendix.

Judges will be evaluating your technical report without your team there to explain it; therefore, **your submission must "speak for itself"**. Please ensure that your main findings are clear and that any visualizations are functionally labeled.

Submissions: Evaluation

The competition will have multiple rounds of evaluation. Your Report will be judged as follows:

- **Non-Technical Executive Summary**
  - *Insightfulness of Conclusions.* What is the question that your team set out to answer, and how did you choose it? Are your conclusions precise and nuanced, as opposed to blanket (over)generalizations?
- **Technical Exposition**
  - *Wrangling & Cleaning Process.* Did you conduct proper quality control and handle common error types? How did you transform the datasets to better use them together? What sorts of feature engineering did you perform? Please describe your process in detail within your Report.
  - *Investigative Depth.* How did you conduct your exploratory data analysis (EDA) process? What other hypothesis tests and ad-hoc studies did you perform, and how did you interpret the results of these? What patterns did you notice, and how did you use these to make subsequent decisions?
  - *Analytical & Modeling Rigor.* What assumptions and choices did you make, and what was your justification for them? How did you perform feature selection? If you built models, how did you analyze their performance, and what shortcomings do they exhibit? If you constructed visualizations and/or conducted statistical tests, what was the motivation behind the particular ones you built, and what do they tell you?

Submissions: Format

Reports can be produced using any tool you prefer (Python Notebook, Shiny Application, Microsoft Office, etc.); however, **your report MUST be in a universally accessible and readable format (HTML, PDF, PPT, Web link)**. It must not require dedicated software to open. For example, if your report is a Python Notebook, it should be exported to HTML. If you create a Shiny App, it should be published at an accessible Web link.

However, please also include the source file used to generate your report. For example, if you submit a PDF with math-type, equations, or symbols, please include your LaTeX source file.

Code should be submitted in a single zipped collection of files separate from your report.

Your team will be sent a Google Form at the beginning of the competition; you will use this form to upload and send in your submitted content. **Submissions MUST be received by 5:00PM HKT (Hong Kong Time) on Sunday, March 28th, 2021. Any submissions received after that time will NOT be evaluated by the judges**.

Tips & Recommendations

This will be a weeklong event; however, you should try to complete as much of your work as possible before the weekend. The extra time may lull you into a false sense of security. Additionally, with your extra time, you should really think about what problem you want to solve. The outcome of this Datathon for you will likely be decided by how well you planned your work.

For data engineering, exploration, and modeling, we highly recommend that you install Jupyter Notebook: http://jupyter.org/install.html. Jupyter Notebook is an interactive, real-time development environment that eliminates many pain points of the standard "terminal + text editor" environment, and is compatible with both Python and R.

We also recommend that your team not try to learn new tools if possible; instead, leverage your existing skills to extract as much insight from the data as you can.

We've compiled 3 additional commonalities of successful teams and 3 pitfalls of unsuccessful teams. Of course, these may not apply to every team, so we recommend that you and your team apply any tips accordingly:

Tips for Success	Try to Avoid
1. Focus on hypothesis testing when brainstorming your research question	**1.** Do not try to exhaust all different models you know just to yield an ideal cross validation accuracy
2. Spend at least 3 hours on your report to ensure strong communication through visualizations and writing	**2.** Do not violate assumptions of statistical models. Sometimes, specific models require specific features so make sure those conditions are sufficient
3. Engage in proper causal analysis. Just because your model passes standard cross-validation checks it does not demonstrate (or even suggest) causality	**3.** Do not pick research statements and blindly stick to it trying to get it to work. Often times, further data exploration will show that it's not even true or worthwhile

Ask for Help

Correlation One's technical product team is here to help. Let us know about your struggles as early on as you can and we may be able to offer advice on how to best move your analysis forward.